Enabling Data Integration in the Rail Industry Using RDF and OWL - the RaCoOn Ontology

Tutcher, Jonathan; Easton, John; Roberts, Clive

DOI:
10.1061/AJRUA6.0000859

Document Version
Peer reviewed version

Citation for published version (Harvard):
Tutcher, J, Easton, J & Roberts, C 2017, 'Enabling Data Integration in the Rail Industry Using RDF and OWL - the RaCoOn Ontology', *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, vol. 3, no. 2. https://doi.org/10.1061/AJRUA6.0000859

ENABLING DATA INTEGRATION IN THE RAIL INDUSTRY USING RDF AND OWL - THE RACOON ONTOLOGY

Jonathan Tutcher[1],

John M. Easton (Corresponding Author) [2],

and Clive Roberts [3]

ABSTRACT

As traditionally infrastructure-centric industries such as the railways deploy ever more complex information systems, data interoperability becomes a challenge that must be overcome in order to facilitate effective decision making and management. In this paper, the authors propose a system based on semantic data modelling techniques to allow integration of information from diverse and heterogeneous sources. The results of work, which aimed to demonstrate how semantic data models can be used in the rail industry, are presented; these include a novel domain ontology for the railways, and a proof-of-concept real time passenger information system based on semantic web technologies. Methods and patterns for creating such ontologies and real world systems are discussed, and ontology-based techniques for integrating data with legacy information systems are shown.

Keywords: Ontology, Linked Data, Railway.

INTRODUCTION

In recent years many railways worldwide have undergone a revival, with growth in passenger numbers driven by factors such as traffic congestion, a desire to work during travel time, and social pressures to switch to "greener" modes of transport. The United Kindom's

[1]Research Fellow, School of Electronic, Electrical and Systems Engineering, University of Birmingham, Edgbaston, Birmingham, B15 2TT. UK.

[2]Lecturer, School of Electronic, Electrical and Systems Engineering, University of Birmingham, Edgbaston, Birmingham, B15 2TT. UK. E-mail: j.m.easton@bham.ac.uk

[3]Professor of Railway Systems, School of Electronic, Electrical and Systems Engineering, University of Birmingham, Edgbaston, Birmingham, B15 2TT. UK.

(UK) railway network is the oldest in the world, and has been steadily growing in popularity; between the years of 1991 and 2011 passenger numbers across the network rose by 67%(Office of Rail Regulation 2011). As demand has increased, the railway industry has invested heavily in infrastructure and rolling stock as it strives to meet the need for greater capacity, while ensuring the railways remain a reliable and safe mode of transport. However, as the cost of new infrastructure has grown and external financial pressures have increased, there is a growing acceptance within the industry that simply "building out of trouble" is no longer a viable long-term solution to passenger growth, and stakeholders are instead seeking to make better use of their existing assets. In many cases this is expected to be achieved through greater use of Information and Communications Technology (ICT), creating a need for accurate, timely information on the state of the infrastructure at the heart of the industries operational and planning decision making processes. In contrast to the infrastructure-centric industry of the past, this new system can be thought of as the *data driven* railway.

The early stages of the industries move towards the data driven railway were characterised by a rush to instrument, monitor and record data on the state of the railway system. Remote Condition Monitoring (RCM) systems, such as Network Rail's Intelligent Infrastructure platform, are now generating huge quantities of data on asset health. However, aside from superficial investigation in the form of thresholding and alarm generation, very little analysis of the corpus gathered is ever carried out. This restricts the amount of new business intelligence that can be gained from the system, and places limits on the overall return on investment. Other, pre-existing ICT systems within the industry are commonly "siloed" - they exist in isolation, with a dedicated set of data collection equipment, databases, and front-ends that serve a specific purpose. The siloed nature of these ICT systems makes it hard to bring together data to answer questions that cross physical system or organisational boundaries, and queries involving data from multiple systems without established interfaces must be carried out manually by human operators. The problem of ICT system fragmenta-

tion in the rail industry is particularly evident in the UK, where privatisation of the railway network in the mid-1990s has led to a complex, multi-stakeholder industry that separates the management of the infrastructure, performed by Network Rail, from the passenger and freight train operators (TOCs and FOCs), rolling stock leasing companies (ROSCOs), and safety and governance authorities including the Rail Safety and Standards Board (RSSB), the Office of Rail and Road (ORR), and the Department for Transport (DfT). It was estimated that in 2011 over 100 shared ICT systems were in use by the UK rail industry(Rail Safety and Standards Board 2012), a figure that does not take into account systems used within individual organisations.

Beyond the railways, other traditionally infrastructure-centric industries faced with the need to make better use of their assets have seen substantial benefits from greater integration of data across ICT system boundaries; a process that enables in-depth, whole-system analyses that can generate new business intelligence. Examples can be found in the upstream oil and gas industries Integrated Information Platform(Sandsmark 2008) and Integrated Operations in the High North projects (Verhelst 2012), and the United States' (US) Capital Facilities Industry, where it was estimated that the adoption of improved information interoperability standards for operations and maintenance could have saved $15.8 billion in 2002 alone (Gallaher et al. 2004).

A key enabler of data integration is the provision of common data models both within the domain of study and linking it to the wider world.

Existing Rail Domain Data Models

A number of research projects and industrial initiatives concerning knowledge management and data modelling for railway data have been undertaken over the last decade, aiming to allow better integration of data between systems. Few have enjoyed significant commercial uptake, although support for RailML (Nash et al. 2004), a cross-industry project establishing comprehensive eXtensible Markup Language (XML) data models for information exchange,

continues to grow in Europe. Other relevant models include efforts by the International Union of Railways (UIC) to develop a new infrastructure model, RailTopoModel, (UIC 2013) and the European Union's (EU) 7th Framework Programme (FP7) InteGRail project (InteGRail 2011), which delivered a basic rail ontology - a semantically richer graph-based representation of domain concepts and relationships that will be discussed in detail in a later section. Many other transportation data models also exist and are widely used; most notably the National Public Transport Access Nodes (NaPTAN) model(Cartwright 2010) and the ArcGIS Esri model(ESRI 2014).

EU interoperability legislation, coming into force across the railway industry from 2015, will also provide incentives for companies to consider novel methods of data management. The EU Register of Infrastructure (RINF) requires that all rail infrastructure operators across Europe provide a basic level of information about their networks(European Railway Agency 2010), a task that has so far caused many companies difficulty. As European interoperability is mandated further, the demand for efficient data management and exchange is likely to grow.

Semantic Data Models

In creating and storing information in a computer system, data context and meaning must be preserved alongside the data itself to allow future re-use. Traditional data storage approaches ensure this by specifying detailed data schemas such that the context of a piece of data can be recalled based on its position in the store, and dealt with correctly. This approach works well for many systems, but makes representation of data not accounted for by the schema very difficult. Interoperability between these systems necessarily requires bespoke interfaces that map data between schemas, as no knowledge of data structure can be deduced or acted on by the computer systems themselves.

Semantic models store data context alongside the data itself in a machine-understandable form, reducing the need for an explicit database schema and allowing greater flexibility in

maintaining and querying the data present. Knowledge is represented by defining a set of entities and building up facts between each entity, each of which has an understood meaning. In this way, an unambiguous representation of each entity and its characteristics is built up without the need for or the constraints of complex data structures. The precise meaning of relationships and entity types within a semantic model are usually defined by an *ontology*: a machine-readable formalisation of how a particular domain or world view works.

The Resource Description Framework (RDF) is a World Wide Web Consortium (W3C) recommendation for the representation of semantic data models that describes entities and the relationships that interlink them as Uniform Resource Identifiers (URIs). Subject-predicate-object facts called *triples* are assembled to build up knowledge graphs, including information about entities and about the semantics of the relationships used. This graph-based representation can be serialized in several ways, including as plain text or XML(Beckett et al. 2013), binary, and inside traditional Structured Query Language (SQL) databases. Examples of RDF triples are shown in Table 1

The RDF itself is technology-agnostic, and thus data is preserved into the future as new technologies and tools are released. The RDF specification only calls for the representation of facts as explained above, and these facts can be serialised in several different ways. Whilst there is still a dependency on the vocabulary and design patterns used when representing data, the meaning of the data is preserved through time, and so its use as a knowledge representation technique fits well with long system lifecycles present in the rail industry.

Ontology

Ontologies are data models that formally describe some problem domain or world view in a machine-interpretable way. By creating ontologies and associating real-world data with them, it becomes possible for computer systems to infer new knowledge in the same way that humans might use "common sense" when considering a set of facts. This inference aids data integration by making data implicit within a model explicit, and simplifies information

systems by reducing the amount of logic required in individual applications when interpreting data.

In practice, two things are required to develop an ontology; a set of controlled vocabularies of terms used within the domain, and a set of related classes and rules that can be used to describe the domain from a particular viewpoint. When using a conceptual data model, developers state the relationship between an item of data and the model of the world, allowing that data to be seen in context by the computer. As an illustration, consider the US English and British English use of the word jelly. Jelly is a perfectly acceptable term in both US and British English, and even has the same syntax (can be used in the same places in sentences), but the meaning of the term - its semantic - is different in the two languages. In British English, jelly is a gelatine-based dessert, whereas in US English jelly is a fruit preserve (a jam in British English). In a conventional data model such as an XML document, the tag jelly can be used ambiguously, because XML schema only enforce the positioning of the tag in the document and the values it can take, not its meaning in that context. Participants in a solely XML-based data exchange could legitimately use the tag jelly for either meaning, even if the designers of the XML schema had a particular usage in mind. By representing the data in a conceptual model this situation can be avoided, because the term jelly is defined as being a type of dessert that is composed of gelatine, and which may have a particular colour, flavour, shape and wobbliness.

Once facts have been entered into an ontology, it becomes possible to use the relationships contained within the model to infer new information about the world; a process known as reasoning. On the simplest level this could involve the user stating that an object "`377 401`" is an instance of "`Class 377`" and the reasoner then inferring that "`377 401`" must be a train because the ontology shows that "`Class 377`" is a type of train. Over and above this type of simple *common sense* reasoning, ontologies can be further enhanced by the additional of rule-based axioms. Rules make it possible for ontologies to capture and use more expressive logical statements that are needed for complex decision making, such as the

following statement:

> "if axle bearing temperature `sensor_x` has a reading of over 100 degrees, then the axle `sensor_x` is monitoring is faulty"

This can be used to infer new facts such as "`axle_bearing_y` is faulty" against the model where relevant. By expressing these rules in the ontology, the operational logic associated with a domain can be stored in the data model, rather than in the code of individual applications, making change management and future development easier (only a single source of domain logic changes, rather than many individual applications).

Ontologies are formalised into machine-readable form by using an *ontology language*; ontology languages provide a defined vocabulary and set of logic with which to build models. By far the most well-supported of these is the Web Ontology Language (OWL). Now in its second major release, OWL is based on *description logic*, and provides a number of sub-languages (called profiles) that trade flexibility with guaranteed computational speed of reasoning. OWL 2 DL is the most expressive of these, allowing complex reasoning across data models at the cost of long worst-case processing times. By contrast, the simpler OWL 2 EL profile guarantees that reasoning will complete in polynomial time with respect to the size of the ontology but limits the range of ideas that can be expressed. A number of other OWL 2 profiles are also specified for other purposes; these include OWL QL, a profile intended to allow standard query languages to utilise ontology (representing relational databases), and OWL RL, which allows expressivity for an ontology to be represented using the logic employed in rule languages.

A CORE ONTOLOGY MODEL FOR THE RAIL DOMAIN

At this stage, the authors would like to introduce the Rail Core Ontology (RaCoOn), an ontology model specifically tailored for use within the railway industry. Although initially

developed with the representation of signalling and rail infrastructure in mind, the model rapidly developed into a general model for the railways, including a "core" of generic railway concepts with extensions capturing particular subdomains (infrastructure, timetabling, rolling stock etc.) and an upper level model to define concepts used more broadly than rail (e.g. transport) The layered design philosophy behind the model is shown in Fig. 1.

Ontology Design and Model Scope

A novel ontology engineering technique based on the NeON methodology(Suárez-Figueroa et al. 2012) was employed in designing the RaCoOn ontologies, based around extracting knowledge from existing railway models and domain experts to inform and validate design decisions. This technique comprised three major steps:

1. Specification: High level requirements were defined, as well as the scope and content specification of system. Several individual ontology modules were defined according to reusability and level of domain detail: an "upper" module for domain-agnostic concepts, a "core" module for railway knowledge, and several subdomain-specific vocabularies including "infrastructure" and "rolling stock".

2. Conceptualisation, formalisation and implementation: Both top-down and re-use oriented approaches were taken in eliciting knowledge for the RaCoOn ontologies, as detailed below.

3. Evaluation and documentation: Ontology modules were evaluated throughout the design process and then validated at the end of the design process.

Specification and Modularisation of Domain Ontologies

Design of application-specific data models is usually driven by a set of functional and non-functional requirements that can be derived from the established needs of the system. Domain models such as the RaCoOn ontologies, however, are intentionally abstracted from any one particular application and are expected to allow representation of concepts without assuming how the data will later be used. The scope of the RaCoOn ontologies was dictated by three initial use cases: an infrastructure visualisation tool, a railway maintenance application, and a signalling design interchange tool. Requirements for these use cases were considered in conjunction with applications and data requirements elicited from recent rail industry data workshops(Roberts et al. 2011), and a high level specification for the RaCoOn ontologies created emphasising commonalities between these use cases.

Conceptualisation and Formalisation of RaCoOn Ontologies

Each ontology module was created by repeatedly iterating over two approaches to model creation: a "top down" method that draws upon expert knowledge to build a hierarchical model of a domain, and a "reuse-oriented" method where existing knowledge was extracted from models such as RailML, Network Rail's Signalling Data Exchange Format (SDEF), and Siemens Rail Automation's Layout Description Language (LDL). In both cases, ontology implementation was performed by defining *ontology design patterns* (ODPs): sets of concepts, relationships, and documentation that define how a particular concept should be encoded in the semantic data model. Fig. 2 shows steps through the iterative process.

The top-down approach aimed to establish a high quality meta-model structure for railway domain knowledge, and to fill gaps in knowledge that may be present when re-using other models. The process performed was as follows:

1. Review scope of initial ontology (or changes for review).

2. Decompose concepts into subcategories, and create competency questions (CQs) around new concepts. For example, when considering a "railway track" entity, a competency question may be: "How can we establish whether a piece of railway track is electrified, and what type of electrification does it provide?"

3. Consider scope of new CQs. A decision on whether they are in or out of scope for the current module is made, and in scope CQs are either implemented or constructed using the reuse-oriented approach.

4. Re-engineer concept into OWL design pattern if appropriate.

The reuse-oriented approach was undertaken to map existing domain knowledge from non-ontological sources into the ontology.:

1. Identify terms for reuse through prompts from previous iterations of this or the top-down process;

2. Analyse term semantics by reviewing documentation and use of a term in the existing model;

3. Re-engineer term into OWL design pattern by either reusing or extending an existing pattern, or creating a new one;

4. Consider new competency questions based on term and design pattern.

Fig. 3 shows decisions made in the creation of an example ontology using this process. New ODPs are shown in the diagram as red stars.

High Level Concepts and Railway Fundamentals

The RaCoOn upper level ontology contains knowledge of generic upper level concepts that transcend the railway domain. Such concepts include space and time, and are mostly reused from existing "gold standard" vocabularies, including:

- The W3C Time Ontology(World Wide Web Consortium 2006), which provides ways of representing instants, intervals, and Allen time relations(Allen 1984). Entities are labelled with start and end times where required, allowing data to be queried based on the time period in which it occurred.

- The W3C Geo(Brickley 2003) and Ordnance Survey (OS) Spatial Relations(Ordnance Survey 2014) ontologies for location positioning.

- The National Aeronautics and Space Administration (NASA) Quantities, Units, Dimensions and Types (QUDT) ontology(Hodgson and Keller 2011) provides an exhaustive list of quantities, units, dimensions and datatypes. These are used in the upper ontology in conjunction with an appropriate design pattern to represent measurements and datatypes.

- ISO15926:2(International Standards Organisation 2003), which provides a meta-model for entity types. The ontology classifies objects into independent (can exist in their

own right), and dependent (existence depends on another entity, such as in the case of a measurement), which is useful in defining acceptable ranges and domains for properties.

The rail core vocabulary ontology is a result of work carried out manually constructing and curating knowledge from other domain models and from UK industry experts. The vocabulary and its sub-modules predominantly draw upon corresponding elements in RailML 2.2, relying on both its XML syntax and human-readable documentation in building an equivalent semantic data model.

THE DELIVERY OF CONSISTENT PASSENGER INFORMATION ACROSS A CHANGING TECHNOLOGICAL LANDSCAPE

In order to demonstrate the feasibility of the use of ontology within the railway industry, the authors joined with Siemens Rail Automation in the UK to produce two technology demonstrators; the work was performed as part of the Future Railway funded "Universal Data Challenge". The first demonstrator, which was presented at the 2014 Institute of Electrical and Electronics Engineers (IEEE) Conference on Big Data(Tutcher 2014), showed how the use of a linked data approach to the handling of asset information could add value as part of a scalable asset management platform. The second demonstrator, which is the subject of this paper, aimed to show how the use of ontology and linked data can help the industry maximise on investment in existing information systems despite changes elsewhere in an increasingly technology-driven railway system. In particular, the demonstrator sets out to show how the use of ontology can provide a bridge between legacy systems and newer replacement services without sacrificing functionality, and how interfaces between such legacy systems and more contemporary linked data-based systems can be set up. As the volumes and variety of data gathered in new information systems on the railway continue to increase,

this demonstrator seeks to illustrate the practical uses of semantic data models in simplifying interfaces and applications, and enriching content.

Train Locator Overview and Key Concepts

Presented as a web application, the Train Locator system demonstrates a number of key areas in which ontologies can allow better integration and management of data in the field of railway Real Time Passenger Information (RTPI) systems. It focuses on the benefits that can be gained by using ontologies to unambiguously describe data to applications, and the ease with which new data in the system can be translated to accommodate existing applications. The following scenarios were demonstrated:

- How two independent RTPI systems can co-exist and share data, without being explicitly designed to do so;
- How new data (train location mileage information) can be quickly integrated into a data model given a new application or physical system upgrade - in this case from track-circuit based location recording to radio-based mileage location recording;
- How ontology reasoning allows a legacy customer information system to continue functioning, even with loss of the initial track circuit location data;
- How ontology rules can be defined in the ontology to provide graceful degradation of functionality in passenger information systems.

In addition to the above, the demonstrator application illustrates the advantages of a linked data-based approach, showing contextualised train station information taken from other

sources, and allowing users to explore information associated with entities such as trains, locations, and schedules.

The storyboard for the demonstrator is as follows:

1. Imagine a railway network equipped with legacy, low resolution train positioning systems, such as track circuits and axle counters. These devices are placed close enough together to drive signalling systems but only provide a low resolution view of where trains are located across a network.

2. The data produced by the train positioning systems is used to (amongst other things) power a number of passenger information systems, including platform boards and third-party applications for mobile devices.

3. As part of an upgrade programme, for example a migration to European Rail Traffic Management System (ERTMS), existing low resolution train positioning equipment on a line is replaced by a more accurate system. Future passenger information systems can be designed to operate using the higher resolution positions from the new system, but existing passenger information systems, that require positional data to be at track circuit level, will all need updating - a costly process that involves many stakeholders if third-party applications are included.

4. In an information landscape utilising ontology, the data being delivered by the positioning systems, and being used by the passenger information systems, is described unambiguously; the computer "knows" exactly what data is available and what is needed by the applications. Rules can be added to the data model describing how data in one form is converted to the other, allowing the system to deliver inferred track circuit-level data to legacy systems based solely on the new, high resolution

location data.

5. By using the combination of ontology, rules, and reasoning, it becomes possible to maintain the functionality of existing applications, despite changes elsewhere in the rail system, without altering the applications' codebase. Ontology will allow the industry to design and implement information systems only once in a changing technological landscape. Old and new applications will be able to co-exist and can be driven by the same underlying data resources.

The demonstrator was designed to showcase the benefits that could be gained through the integration of data across a simple semantic data model using only a few very simple rules and ontological axioms. In order to achieve this two simple passenger information applications were created, and ontology reasoning was used to remove these applications' reliance on specific input data types.

The demonstrator itself, available at http://purl.org/rail/trainlocator, is a website that provides a number of views to simulate real world railway customer information systems. Each view illustrates a usage scenario, and the application is designed to allow users to understand the effects and advantages of differing ontology constructs on the system. Train movement data is provided by simulated values, which update the website in real time and drive outputs on each page.

The key technological components used in the presentation of this demonstrator are:

- Stardog, an RDF triple store used to store all data (ontology and resources). Stardog is a scalable off-the-shelf product that provides several levels of ontology reasoning, from the schema reasoning described above, to the ability to read custom-written rules. It conforms to W3C standards on linked data storage and presentation, allow-

ing a generic interface between the application and data store to be created;

- The train movement simulator, residing on the web server, which updates the locations of a set of trains as they pass through the demonstrator's railway network. Train locations are simulated through internal logic and pushed to the Stardog server through its linked data endpoint. Controls on the demonstrator website allow the user control over whether the simulator sends legacy (track circuit) or high resolution (mileage) train position data;

- A web user interface, written using modern web technologies - Hypertext Markup Language (HTML), Cascading Style Sheets (CSS), and Javascript. This front end provides all of the application functionality, and queries the Stardog data store directly for each function. Logic in the web front end is limited purely to presentation details; all other information about interactions between trains, infrastructure, and location is stored and computed in the triplestore.

These three components communicate via the SPARQL 1.1(Prud'hommeaux et al. 2008) linked data protocol, and data is exchanged in linked data at all points. Further input and output applications could quickly be realised by leveraging ICT industry standard practices for linked data and concepts shown in the core railway ontology.

The web user interface shows information in any one of three scenarios:

- Legacy Departure Board System (Using Track Circuit Data). In this scenario, a user can select a train station and view a very basic simulation of a platform-based passenger information board, including departure point, destination location, scheduled, and expected times. Expected times are calculated based on the position of trains

on a track circuit (such as would be provided by a train describer system), which is queried directly from the triple store. The current track circuit of each train can also be displayed for exploratory purposes;

- Train Position Map (Using Mileage Data). The train position map shows the "live" locations of each train on the network. The system queries the ontology for mileage location, and displays it in line with the train's route through the network. Through rule reasoning, the ontology provides the train position map with the most relevant data should both be available;

- Entity Information View (Using Linked Data & Inference). The final view is provided should a user want more information on a particular train, station, or location. The application requests information from the ontology about the location in question, and returns useful information. In the case of train services, inference provides information about the rolling stock itself as well as the train service; for locations, reasoning provides additional information such as touching/neighbouring entities and line reference information.

A summary of the behaviour of the ontology given differing applications and input data is shown in Table 2.

Design Patterns and Reasoning Devices

Infrastructure and Location Storage Design Pattern

Infrastructure & location data is stored in the train locator demonstration model as linked data, following patterns defined by the core ontology. Data taken from ATOC working timetable files was used as a base for modelling train movements, and track circuits

were added manually, using simulated track circuit distances. Each "Track Circuit" object has a start location and end location, each of which have an associated mileage and Global Positioning System (GPS) co-ordinates, and these track circuits are aggregated into ":ServiceNode" objects that are referenced in timetable data. Fig. 4 shows an example Service Node associated with a track circuit, which is in turn associated with maximum and minimum locations at points along the track infrastructure.

By linking track circuits to mileages and known pieces of infrastructure, inference can provide train services associated with them with further information. For example, in the case of a train stoppage or cancellation, passengers using linked-data based applications could check the next station's facilities and connections based on the train they are currently on, although this functionality is not shown in the demonstrator.

Reasoning to Allow Legacy System Functionality Given New System Input Data

In order to provide legacy system functionality when a system upgrade occurs, a rule is constructed and added to the triple store. Rules are custom-based reasoning patterns that a triplestore applies to matching data at query-time. The aim is to capture the following knowledge:

> "If a train's current mileage is between the minimum and maximum mileages of a particular track section, and on the same line, the train is defined as being in that track section"

When encoded as a SPARQL rule, this logic leads the reasoner to perform the following actions:

1. Check for current node's line reference;

2. Filter list of possible track circuits to only those on current line;

3. Retrieve minimum and maximum mileages for each candidate match;

4. Identify track circuits with mileages within range of current train's mileage;

5. Assert that the current node is associated with the matching track circuit.

Consequently, whenever a legacy application now requests a node's track circuit location, this rule is checked and the correct track circuit returned whether it was encoded explicitly by an input system, or calculated based on a train's current mileage position.

Reasoning to Allow Improved Resilience of Information Systems during Degraded Service

The strengths of an ontology-driven data store do not only allow the mapping of new data back into other forms for use in legacy systems, but also make it possible to increase data availability during periods of degraded system reliability. Using the capability of the system to interlink data, a hierarchy of "preferred" properties were specified for each system concept, and these hierarchies used with closed world rule reasoning to find the best available data for a particular application. Recall the following scenario from the storyboard presented earlier in this paper:

- A railway line has recently been upgraded to ERTMS operation, and now provides very rich location information for each train on the track, rather than only track circuit occupation details;

- New applications for customer information and service monitoring are built using the new, more accurate ERTMS location information. It is desirable, however, for these systems to continue functioning in times of degraded operations - for instance if ERTMS systems are unavailable and the line reverts to fixed block operation.

In this case, the usual approach would be to include application logic to search for available systems and make a decision specified at system design time as to which data source to choose - an approach which is inflexible and unsustainable in a complex system.
To enable the data model to find which data to provide for a train location application, the following pattern encodes knowledge of "preferred systems" (see Fig. 5). This shows several OWL classes (marked with yellow circles) related to each other through RaCoOn properties (marked on arrows) forming transitive ":preferredOver" relations. Thus, a reasoner can infer that an "is:CrsLocation" instance is preferred to a "vocab:RailwayMileage", and can choose to prioritise data of this type.
With this knowledge of which system of measurements is preferred given data availability, it is now possible to encode a rule that states:

> "If entity X has multiple locations associated with it, and one is preferred (location Y) over the other (location Z), then insert a new fact: entity X $->$ preferredLocation $->$ location"

As a result of the inclusion of these rules, systems utilising the :preferredLocation property will automatically be presented with the most accurate data for their needs. *It is important to note that applications have varying requirements for location data (some rely on GPS co-ordinates, others rely on Computer Reservation System (CRS) codes, and others on data with other constraints). The pattern above does not ignore these constraints; they are repre-*

sented through other clauses in the query.

Implemented System Demonstrator

The demonstrator web application includes several views which show the effect of reasoning based on location, as discussed above. These views are described in the following sections.

Admin Page: Scenario Control

The Train Mapper home page, accessed when the user first contacts the system, briefly explains the aims of the demonstration and gives users control of the various scenario configuration options. These options influence the behaviour of both the legacy "Departure Boards" view, and the "Map View" application. On-screen controls allow users to select the data supplied to the system by the simulator: either track circuit data, mileage-based position data, or both. Further configuration options turn reasoning on and off within the web application, allowing users to see the effect with or without rules being triggered in the ontology.

Legacy Departure Boards View

The departure boards view (see Fig. 6) shows trains soon to arrive and depart from a station. These are determined by querying the triplestore for relevant services with an appropriate arrival time, and station information if present. Expected train times are naively obtained through adding a `:trainTime` property to every track circuit, and calculating the difference in this property's at the current train's location and the station being viewed.
If the "Track circuit data" data source is turned on, the departure boards view utilises no ontology reasoning whatsoever. Instead, it is presented as a legacy system using linked data as a data storage and interchange format. There are advantages even to this approach, as can be proven by the success and uptake of the Linked Open Data movement on the World

Wide Web. If the "track circuit data" data source is missing, however, ontology reasoning steps in, and resolves live train locations to track circuits for the benefit of this application.

Map View: Dynamic Train Progress

The dynamic train progress page (see Fig. 7) allows a user to track the progress of a train in real time, using mileage values resolved from a fictitious moving block signalling system. Users can select the train they want to track, and watch its position change across the map.

- With only the "mileage data" source turned on, this display uses no inference and displays the current mileage of the train selected on a map.

- With both mileage data and track circuit data, this display calls the ontology to ascertain the priority of these location values (as described above), and displays the mileage location, with its track circuit displayed as a secondary information source.

With only track circuit data available, the ontology resolves a less accurate position for the train based on available information. Whilst it would have been possible to build this logic into the application itself, this approach quickly becomes complicated and hard to maintain when deployed as part of a more complex system.

Map View: Track Circuit Information

Finally, the track circuit and entity views (see Fig. 8) allow users to view more detailed information about each track circuit, or other entity. With reasoning disabled, queries used to populate this view bring back only explicit information held in the infrastructure database about track circuit information. However, with reasoning enabled, links between track circuit locations and other infrastructure items become apparent, and users are able to browse

information about train stations, maintainers, and nearby trains. This view is included to further illustrate the use of ontology reasoning to enrich knowledge and convey useful inferred information.

Alternative Use Cases

The demonstrator discussed in this paper was designed to illustrate how the application of ontology, developed according to a modular approach and using a set of basic design patterns, can deliver large potential benefits when used to integrate multiple industrial information systems. The benefits described in the scenario (e.g. selection of the most appropriate form of data based on the task being performed and the resources available, the delivery of inferred results from queries, or the reusability of the RDF data resources) also apply to a wide range of other industrial use cases, a few examples of which are given below.

Railway Operations Management and Train Routing During Degraded Railway Service

Whilst an ontology will not in itself evaluate decisions on train routing, the ability to provide data at the most appropriate level of granularity that is available can inform human signallers and computer algorithms and help them to make operational decisions based on the most accurate information available at the time. For example, consider a scenario in which two trains are waiting outside a major interchange station, and one unoccupied platform is available. At present, selecting which of the trains can proceed and which should be delayed depends mostly on a controller's intuition to work effectively; this relies on highly experienced individuals fulfilling the same role over months or years and presents challenges in terms of the retention of corporate knowledge. Using a more integrated data approach, where data is drawn from a number of ICT systems as needed, would allow a more informed decision-making process, particularly for staff members transferring into the geographical area under control. Ontology reasoning could infer typical train capacity in absence of it being known, or show actual capacity if it is. Likewise, information on connections from the

following stations could be displayed if known, or based on a rule if not.

Railway Maintenance on Tracked and Untracked Rolling Stock Assets

Using the same approach, plus knowledge of rolling stock composition (as provided by the infrastructure ontology), rolling stock maintainers can be informed of likely asset failures in absence of monitoring information. If a particular class of railway vehicle is known to develop a fault after a certain number of miles, it is possible for the ontology to display these likely faults on appropriate vehicles, and not to display them on vehicles with more detailed explicit information stored.

Cross-Railway Train Position Reconciliation

The property translation pattern used to map mileage values into track circuit values is only one example of the ability of semantic data models to accommodate transition from legacy systems. Whilst ontologies cannot themselves provide very complex algebraic mappings from new systems to old (for example, geographic transforms), reasoning allows more common sense properties to be conveyed between systems with very little overhead. An example of this may be in resolving a problem encountered by open data enthusiasts when reconciling London Underground and Network Rail train movement data in regions where the two providers overlap. Where multiple systems log the same information about trains in different ways, ontology rules and mappings can help to align data to be appear coherent. As these system interactions change, rules can be updated, and no change to application code is needed.

Potential Benefits of Implementation

Industry-wide ontology models for rail have been the subject of significant discussion in the UK rail sector in recent years. The 2013 Network Rail Technical Strategy(Network Rail

Limited 2013), which outlines the UK Infrastructure Manager's priorities for investment in new technology over the period 2014 - 2019 and beyond, suggests that developing the research into ontologies for rail to an "implementation ready" level (i.e. Technology Readiness Levels 5 - 7) would cost the industry up to £1 million; however, as is often the case in this area the document presents no indication of the value of the potential benefits. An estimate for the financial benefits resulting from the implementation of ontology in the UK rail industry can be found by reference to other domains. As previously mentioned in this paper, the US National Institute of Science and Technology's cost analysis of inadequate interoperability in the capital facilities industry(Gallaher et al. 2004) found that $15.8 billion could have been saved in 2002 through improved information interoperability, a figure that represents between 1% and 2% of revenue for that year. The capital services industry, which deals with the construction and management of large commercial and industrial facilities, is similar the the UK rail industry in many respects; it consists of a large number of stakeholder organisations, each with their own ICT provision, which specialise in delivering infrastructure with a long lifecycle - as a result of this, the industry is an appropriate analogue to the railways. On this basis, taking the 1% to 2% revenue figure for the capital services industry and translating it into the UK rail industry, where the Train Operating Companies received fare revenues of £8.2 billion from passengers in the year 2013/2014(Office of Rail Regulation 2015), results in between £82 million and £164 million of potential savings annually. If only a very small proportion of this figure were to be realised in practice by the industry through the use of a common data model such as the RaCoOn ontology, then the financial benefits would be very significant.

The design patterns and processes demonstrated in this paper have diverse applications across the railway; in particular, the demonstrator highlights a fundamental technique (the ability to utilise the most appropriate available resource of a given type) that can be implemented wherever multiple real-world systems provide the same type of information into a data store. Across the industry the ability to automatically select the most appropriate

information resources for the selection available means that legacy software packages can still function in environments using upgraded information stores, maximising the useful lifetime and return on investment from these software packages. Furthermore, by moving the data dependency to the models and data repositories, rather than the applications, adopting the proposed design patterns will enable data-centric, rather than application-centric, management of the selection of appropriate information; reducing the complexity and cost of implementing business logic changes in software.

Limitations of Approach

The adoptation of common semantic models, such as RaCoOn, have many potential benefits to offer the railway industry. Care must be taken however, to avoid thinking of the technology as a 'silver bullet' that will fit perfectly into every possible data exchange scenario. In enterprise contexts, OWL/RDF systems offer a pragmatic solution to the representation of domain semantics, however, there are limitations to the current implementation technologies as outlined in the following sections.

Scalability, Reasoning Performance, and Expressivity

Many of the benefits of using ontological models in information systems arise from their ability to infer new knowledge from existing data. Whilst some of this inference can be done in an efficient manner, much of the OWL DL language requires reasoning algorithms that do not scale to large volumes of data. A trade-off between reasoning performance and scalability is required, which currently prohibit many useful axioms being used in large applications. Ongoing research in 'web-scale' reasoning techniques combined with state-of-the-art RDF graph storage technology is likely to bring increased performance in the future, but applying reasoning techniques over large datasets, represented using highly expressive OWL models, is currently a significant technical challenge.

Architecture and Distribution

Cross-enterprise data exchange is necessarily decentralised, and requires transmission and consumption of information between many systems and parties. While ontological models make it easy to refer to the same concepts universally, they do not address the practicalities of actually publishing and consuming information. Data sharing on the wider semantic web shares this issue: to make use of another dataset, one must either download it in its entirety to interact with locally, or rely on the data provider's processing power and availability using a SPARQL endpoint. Possible alternative architectural approaches to the data provisioning issue include the use of bespoke Service Oriented Architectures, and Linked Data Fragments, which use small, targeted data dumps to facilitate local querying of federated data. However, work remains to be done in this area before a 'gold standard' architectural template can emerge.

Versioning and Change Control

Although ontological models afford a great deal of flexibility and backwards compatibility in their design and modification, it is still possible for changes to, or removal of, existing axioms from published ontologies to result in incompatibilities between systems. Web ontologies often present version-specific Internationalized Resource Identifiers (IRIs), so-called 'Version IRIs' in addition to canonical IRIs to allow for users that wish to fix their application on one version of an ontology, but the problem of change management through a network of ontologies has yet to be formally addressed.

CONCLUSIONS

As demonstrated by the UK Rail Technical Strategy, which states that "Common architectures and protocols would facilitate integration and information-sharing. Costs would be lowered and services improved"(Rail Safety and Standards Board 2012), there is an appetite for greater data integration within the rail industry. This paper discusses the application of

ontology and semantic data modelling techniques to provide better knowledge management across large complex systems as one possible response to that aspiration. Development of technologies for use in the Semantic Web has led to creation of mature toolsets for creation of computer-understandable domain ontologies, as well as for reliable storage, querying, and data exchange of semantic models. These technologies have been proven in use on the web, and can now be exploited to provide ways of sharing enterprise data across industries such as the railway.

A proof-of-concept demonstrator was presented, drawing upon a novel rail domain ontology created at the University of Birmingham. The Train Locator application implemented a use case in which ontologies and semantic web technology were used to contextualise and enrich information from multiple systems. Simulating two different data sources, the demonstration showed how domain models allow presentation of data independently of its original syntax, such that two applications could be driven by data not originally intended for that use. The patterns for data usage across systems outlined in this document are transferable to other application domains, both within the rail and in other similar industrial settings. The application's implementation using off-the-shelf and open source components, including standard web technology stacks, shows the ease with which such applications can be built. Future work, also being funded by Future Railway in the UK, will focusses on standardising methods for collaborative ontology creation in the rail domain with a view towards encouraging uptake of such models across the industry.

ACKNOWLEDGEMENTS

The authors would like to thank Siemens Rail Automation UK, as partners in the Future Railway Universal Data Challenge project. The work on railway core ontologies was undertaken as a PhD CASE Studentship, funded by the Engineering and Physical Sciences Research Council.

APPENDIX I. REFERENCES

Allen, J. F. (1984). "Towards a general theory of action and time." *Artif. Intell.*, 23(2), 123–154.

Beckett, D., Berners-Lee, T., Prud'Hommeaux, E., and Carothers, G. (2013). "Turtle - Terse RDF Triple Language - W3C Candidate Recommendation, <http://www.w3.org/TR/turtle/>.

Brickley, D. (2003). "W3C Basic Geo Vocabulary, <http://www.w3.org/2003/01/geo/>.

Cartwright, M. (2010). "NPTG and NaPTAN Schema Guide v2.4." *Report no.*, Department for Transport, <http://www.dft.gov.uk/naptan/schema/2.4/doc/NaPTANSchemaGuide-2.4-v0.57.pdf>.

ESRI (2014). "Transportation Data Model - GIS for Transportation, <http://www.esri.com/industries/transport/community/data_model>.

European Railway Agency (2010). "Rail System Register Of Infrastructure - Final Report." *Report no.*, European Railway Agency, <http://www.era.europa.eu/Document-Register/Documents/IU-Recommendation on specification of RINF-Final Report.pdf>.

Gallaher, M., OConnor, A., Dettbarn, J., and Gilday, L. (2004). "Cost Analysis of Inadequate Interoperability in the U.S. Capital Facilities Industry. Available online from http://iringtoday.com/wordpress/wp-content/uploads/2012/02/NIST-Interoperability-Report.pdf. Last accessed 4th October 2015.

Hodgson, R. and Keller, P. J. (2011). "QUDT - Quantities, Units, Dimensions and Types, <http://www.qudt.org/>.

InteGRail (2011). "InteGRail - Intelligent Integration of Railway Systems, <http://www.integrail.info/>.

International Standards Organisation (2003). "ISO 15926-2:2003 Integration of life-cycle data for process plants including oil and gas production facilities." *Report no.*, International Standards Organisation.

Nash, A., Huerlimann, D., Schuette, J., and Krauss, V. (2004). "RailML-a standard data

interface for railroad applications." *Publ. WIT Press*.

Network Rail Limited (2013). "Technical Strategy: A Future Driven by Innovation. Available online from http://www.networkrail.co.uk/publications/technical-strategy.pdf. Last accessed 4th October 2015.

Office of Rail Regulation (2011). "National Rail Trends 2010-2011 Yearbook Data, <http://www.rail-reg.gov.uk/upload/xls/ nrt-yearbook-2010-11.xls>.

Office of Rail Regulation (2015). "GB Rail Industry Financial Information 2013-14. Available online from http://orr.gov.uk/__data/assets/pdf_file/0005/16997/gb-rail-industry-financials-2013-14.pdf. Last accessed 4th October 2015.

Ordnance Survey (2014). "Spatial Relations Ontology, <http://data.ordnancesurvey.co.uk/ontology/spatialrelations/>.

Prud'hommeaux, E., Seaborne, A., and Seabourne, A. (2008). "SPARQL Query Language for RDF, <http://www.w3.org/TR/rdf-sparql-query/>.

Rail Safety and Standards Board (2012). "The Railway Technical Strategy 2012." *Report no.*, Rail Safety and Standards Board, <http://futurerailway.org/RTS/>.

Roberts, C., Easton, J., Davies, R., Sharples, S., and Golightly, D. (2011). "Rail Research UK Feasibility Account: The Specification of a System-wide Data Framework for the Railway Industry Final Report. Available online from http://p.sparkrail.org/record.asp?q=PB022964, last accessed 6th October 2015.

Sandsmark, N. (2008). "Integrated information platform for reservoir and subsea production system (rev 2), <https://www.posccaesar.org/raw-attachment/wiki/IIP/IIP_sluttrapport_2008_Public.pdf> (October).

Suárez-Figueroa, M. C., Gómez-Pérez, A., and Fernández-López, M. (2012). "The NeON Methodology For Ontology Engineering." *Ontol. Eng. a Networked World*, M. C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, and A. Gangemi, eds., Springer Berlin Heidelberg, Chapter 1, 9–34.

Tutcher, J. (2014). "Ontology-driven data integration for railway asset monitoring applica-

tions." *IEEE International Conference on Big Data*, 85–95 (Oct).

UIC (2013). "RailTopoModel - Railway Network Description, <http://railml.org/tl_files/railML.org/documents/science/201213_UIC_RailTopoModel_DraftDec13.pdf

Verhelst, F. (2012). "Integrated operations in the high north: Final report (rev 2), <http://iringtoday.com/wordpress/wp-content/uploads/2012/02/NIST-Interoperability-Report.pdf> (June).

World Wide Web Consortium (2006). "Time Ontology in OWL, <http://www.w3.org/TR/owl-time/>.

List of Tables

TABLE 1. Table Showing Examples of RDF Triples

Subject	Predicate	Object
:Pendolino390003	rdf:type	:Train
:Pendolino390003	:operatedBy	:VirginTrains
:Pendolino390003	:location	:CoventryStation
:CoventryStation	rdf:type	:TrainStation
:CoventryStation	:operatedBy	:VirginTrains

TABLE 2. Information sources for train locator application scenarios.

Application Scenario / RTPI System Type	Track Circuit Data	Mileage (Moving Block) Data
Legacy Departure Board System	Asserted (real) track circuit data	Inferred track circuit based on train mileage.
Train Position Map	Inferred (approximate) train location based on known track circuit positions.	Asserted (real) mileage data.
Train Position Map (When Both Sets of Location Data are Available)	Rule reasoning chooses optimum location object for the task.	

List of Figures

FIG. 1. Layered design philosophy underpinning the RaCoOn model.

FIG. 2. Block diagram showing ontology design process

FIG. 3. Example competency questions and paths to ontology creation

FIG. 4. Ontology graph showing track circuit positioning.

FIG. 5. Ontology graph showing the "preferredOver" relation between locations.

FIG. 6. Train locator departure boards view. Map Data ©OpenStreetMap contributors, CC-BY-SA, Imagery ©Mapbox.

FIG. 7. Live train information map view in train locator. Map Data ©OpenStreetMap contributors, CC-BY-SA, Imagery ©Mapbox.

FIG. 8. Track circuit detail and track circuit boundary overview screenshot. Map Data ©OpenStreetMap contributors, CC-BY-SA, Imagery ©Mapbox.